From Genotype to Phenotype: Assembly and Annotation of Two Fungal Genomes

Brianna Barry

5 May 2017

In partial fulfillment of the requirements of graduation with

Special Departmental Honors in Biochemistry

Thesis and research supervised by Dr. Christine Hawkes

Associate Professor

Department of Integrative Biology

University of Texas at Austin

Acknowledgements

Funds for DNA sequencing and computational resources were provided by an Ecolab grant awarded to Dr. Christine Hawkes.

The Texas Advanced Computing Center (TACC) at the University of Texas at Austin for providing provided vial computing resources that significantly contributed to the research results reported within this paper.

Dr. Christine Hawkes has served as my PI and mentor for nearly three years. I owe most of my research abilities to her. She encouraged me to pursue a thesis project, and she generously opened her lab to me for the endeavor. Dr. Hawkes has inspired my work and dedication to research, and she has been the most influential professor in my undergraduate education.

Dr. Elise Worchel Connor, a recently graduated PhD student in the Hawkes lab, aided the DNA extraction and sequencing phases of the project, and she tirelessly worked through programming errors with me during the genome assembly and assessment phases. Elise also provided guidance on the transcriptomic analysis.

Dr. Hannah Giauque helped me select the two fungi of interest, and she generously provided the transcriptome data from her thesis work to aid the annotation phase of the project.

Table of Contents

Abstract

All organisms in an environment interact with that environment, including with both other organisms as well as the abiotic surroundings. Symbiosis is defined as a close, long-term association between organisms of different species that commonly results in novel structures and/or metabolism. Ascomycetes, also known as sac fungi, are the largest phylum of Fungi with over 64,000 currently known species. Within this diverse group, beneficial and detrimental associations with specific plant hosts are observed. I selected an antagonist, *Cochliobolus kusanoi*, and a mutualist, *Penicillium pinophilum*, for our analyses. These endophytes were isolated from *Panicum virgatum*. I conducted paired-end sequencing with an Illumina HiSeq 4000 system to investigate the genetic underpinnings of such complex relationships. Genome assembly was completed with several programs for performance comparison, namely Velvet, MaSuRCA, and SOAPdenovo2. The de novo assemblies were assessed for completeness with BUSCO and QUAST. The final phase of the project is annotating the two genomes by following the standard DOE-JGI Fungal Genome Annotation Pipeline recommendations and incorporating our previously collected transcriptome data for these fungi to improve the annotation. Genes related to the fungal phenotypes will be predicted and functionally annotated, and comparative analysis will enable the visualization of specific gene structures and domain compositions. I successfully created a unique and accurate workflow for future fungal genomics research beginning with DNA extraction through genome annotation.

Introduction

Background

In an era with increasingly affordable genome sequencing technologies, genome assembly and annotation have become projects for individuals, not research groups or large collaborations. Thus, there are numerous computational tools available to assemble, assess, and annotate a genome of interest. With an estimated 5.1 million species, fungi represent one of the largest branches of the Tree of Life. Fungal genomics provides insight into both genetics and applied genetics, including mechanisms of fungal genome evolution, fungi-specific gene family innovations, genomic potential for sexual cycles, functional genomics, structural and functional characterization of protein-coding and non-coding genes, and other genomic features. As genome annotation has become more popular among individuals, it has also become more challenging. Shorter read lengths from second-generation sequencing platforms make assembly more difficult by decreasing the contiguity[1]. Novel genes in recently sequenced genomes can be difficult to identify, and the annotation data sets still need to be updated and merged. I aimed to use two fungal genomes to identify fungal genetic traits associated with mutualist or antagonist phenotypes; to this end, I first determined the best approach for fungal genome assembly.

All terrestrial plants measured to date form symbiotic relationships with fungi. I focused on foliar fungal endophytes, one group of fungal symbionts that live within plant leaves for at least part of the fungal life cycle without evoking symptoms of harm from the plant hosts[2]. These widespread fungal-plant associations can influence host fitness, plant community composition, soil nutrient availability, and more. Foliar fungal endophytes are particularly well known for moderating plant stress responses[3]. These endophytes can confer plant drought resistance by strategically avoiding drought via increased water uptake or decreased transpiration rate or by tolerating drought through osmotic adjustment[4,5,6,7,8]. Previous grasslands research demonstrated that fungal symbionts differentially interact with grass hosts based on both biotic and abiotic conditions, and endophytes sort by environment across a local precipitation gradient[9,10]. However, endophyte symbioses are not predictable from local environmental conditions, and endophyte community composition is not explained by plant host traits, spatial factors, or vegetation structure[10,11]. Endophytes differentially confer plant trait plasticity in a taxon-dependent manner[10]. While plant-fungal symbioses affect plant drought response, predictive frameworks endophyte effects cannot be derived from community composition data alone. Individual fungal taxa must be understood from a genotypic level to inform functional models.

I sequenced the genomes of two endophytes in the Ascomycota, *Cochliobolus kusanoi* and *Penicillium pinophilum*. The genus *Cochliobolus* contains 55 known species, many of which are destructive plant pathogens that affect agricultural yields. This genus forms a complex with *Bipolaris* and *Curvularia*, which commonly contain grass pathogens with a worldwide distribution[12]. *Cochliobolus kusanoi* is an antagonist to *Panicum virgatum*, the plant host studied here. *C. kusanoi* produces secondary metabolites with antimicrobial, antioxidant, and cytotoxic activities[4].

Penicillium species are common fungi living in a diverse range of environments and are most known for the penicillin-producing taxa. *Penicillium* species, commonly pathogenic, generally decompose organic materials and cause rotting in food products by producing mycotoxins. The genus contains 354 species, and these fungi can be recognized by their dense, brush-like spore-bearing structures called penicilli. *Penicillum pinophilum* acts as a mutualist within *P. virgatum*.

Assembly Approaches

The Human Genome Project, completed in 2003, utilized the hierarchical shotgun approach to sequence the genome. Following the draft assembly, genomes were assembled with Sanger sequencing, which produces read lengths of over 800 base pairs (bp). Second-generation sequencing technologies like those offered by Illumina are short read methods, ranging from 50 – 400 bp. Short-read sequencing is significantly less expensive, and new assembly methods have been developed to accommodate the shorter sequence lengths. However, higher coverage, or read repetition, is required to produce long enough sets of overlapping sequences to form consensus regions. These regions, known as contigs, can be accurately assembled into larger scaffolds that incorporate gaps into the sequence. Most assemblies based on short reads are only draft quality, containing significant gap regions and errors. Thus, the selection of a genome assembly program is critical to maximizing the information obtained from a sequencing method and to optimizing the sequencer-to-assembler match to produce the best possible assembly.

Assembly programs generally use one of two approaches: the overlap-layout-consensus (OLC) assembly or the de Bruijn graph assembly. The OLC assembly method begins by attempting to compute all pairwise overlaps between reads using sequence similarity. Then, the algorithm produces an alignment, or layout, of all overlapping reads. From this layout, a consensus sequence is chosen by scanning the multiread alignment column by column[13]. Most assemblers for Sanger sequencing are based on the OLC approach. OLC assemblers are flexible with read lengths and robust with sequencing errors. However, second-generation sequencing technologies produce short reads and nonuniform coverage, which presents a challenge for this type of approach. To use an OLC assembler with short reads, sequence coverage must be high to produce even a draft assembly.

The de Bruijn graph method eliminates pairwise overlap computation, making it ideal for sequencing platforms like Illumina. Pevzner et al. designed this method for the assembly of Sanger reads using the Euler assembler[14]. However, the approach has been widely applied to second-generation sequencing techniques. The foundation of the technique lies with the generation of the de Bruijn graph. This is an efficient way to represent a sequence in terms of its k-mer components. First, a graph of length k is split into its k-mer components, and these substrings are then assigned to a directed edge in a graph connecting nodes A and B. These nodes connect pairs of k-mers with overlaps between the first k-1 nucleotides and last k-nucleotides. Any direction, or Eulerian path, through the graph that visits every edge (first or last k-1 nucleotides) exactly once forms a draft assembly of reads. These graphs become very complex with many possible Eulerian paths and intersecting cycles. Thus, to deduce an accurate assembly, mate-pair or reference sequence information should be interpreted alongside the graph. Additionally, reads should be retained to help disentangle the possible graphs since k-mers provide less information than raw reads.

Assembly Programs Selected for Testing

Based on my Illumina 2x150 HiSeq 4000 paired-end sequencing data, the de Bruijn graph approach was selected for assembly. Multiple programs use this approach, including Velvet, ALLPATHS-LG, EULER-SR, and ABySS. Three de Bruijn assembly programs were chosen for comparison in this study, namely Velvet, SOAPdenovo2, and MaSuRCA.

Velvet is a short-read assembler designed to manipulate de Bruijn graphs to eliminate errors and resolve repeats in a two-step process after construction and simplification[15]. Velvet possesses an error correction algorithm that merges sequences that belong together. Then, the repeat-solving algorithm separates paths sharing local overlaps. Velvet's performance was originally tested on four different reference genomes: *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens*[15]. When tested on a bacterial species, Velvet demonstrated a higher N50 values and a lower average error rate compared to two other short read assemblers, SSAKE and VCAKE. Velvet required slightly more memory but significantly less run time. Velvet had higher sequence coverage that SSAKE but lower coverage than VCAKE. The Velvet developers recommend resolving small repeats with additional paired read information beyond simply using Velvet.

SOAPdenovo2 builds on SOAPdenovo, which is designed to assemble genomes *de novo* (i.e., without a reference genome) using next-generation sequencing short reads[16]. SOAPdenovo2 boasts a new algorithm design that reduces memory consumption in de Bruijn graph construction. Compared to its predecessor, SOAPdenovo2 also resolves more repeat regions, increases coverage and scaffold length, improves gap closing, and optimizes for large genomes. Like SOAPdenovo, SOAPdenovo2 possesses six modules for read error correction, de Bruijn graph construction, contig assembly, paired-end reads mapping, scaffold construction, and gap closure. SOAPdenovo2 was tested on the Assemblathon1 benchmark dataset, the YH Asian Genome, two bacterial species, and the common eastern bumblebee. SOAPdenovo2 showed marginal declines compared to ALLPATHS-LG in most quality metrics except for the YH Asian Genome assembly, but the memory and run time required were significantly reduced in all tests. Thus, SOAPdenovo2 is recommended for *de novo* genome assembly for eukaryotic genomes.

MaSuRCA, Maryland Super-Read Celera Assembler, uses a hybrid approach to short-read assembly[13]. MaSuRCA has the computational efficiency of de Bruijn graph methods and the flexibility of overlap-based assembly strategies, allowing for variable read lengths and tolerating sequencing error. MaSuRCA generates "super-reads" by transforming large numbers of paired-end reads into smaller, longer reads. Consequently, combinations of Illumina reads with different lengths can be assembled together with longer reads from other sequencing platforms like 454 and Pacbio. MaSuRCA was originally tested against ALLPATHS-LG and SOAPdenovo2 using a bacterial species and chromosome 16 of the mouse genome[13]. For the best assembly, the developers of MaSuRCA recommend supplementing the original data with long reads. Without long read data, MaSuRCA performed on par with ALLPATHS-LG and significantly better than SOAPdenovo2 on the bacterial species, and MaSuRCA significantly outperformed the others when long reads were incorporated. Similar results were obtained using the mouse chromosome. MaSuRCA has subsequently been used to assembly tree, cow, macaque, buffalo, cat, tarsier, ant, and fly genomes[13].

Scaffold_Builder is designed to improve genome assemblies by merging *de novo* genome assembly with a reference genome[23]. Scaffold_Builder generates scaffolds based on similarity to a closely related reference sequence, independent of mate-pair information. Thus, the program is designed to complement *de novo* assemblies generated with other programs. Scaffold_Builder was tested on several bacterial species and was shown to decrease the number of contig sequences by 53% while doubling their average length. The primary limitation of this program is the availability of a quality, closely related reference genome. Scaffold_Builder can be used to increase the completeness of assembled genomes. For example, if SOAPdenovo2 was used to assemble a genome for annotation, Scaffold_Builder would likely be necessary to enhance the genome quality for downstream analyses.

Trimming and K-mer Correction

The de Bruijn graph assembly method cuts reads into substrings of length k, called k-mers[17]. De Bruijn nodes are known as (k – 1)-mers, and edges are the k-mers from the reads. The choice of k-mer length is critical for assembly. Repeats longer than k nucleotides can cause short contigs by tangling the graph. However, a large value of k increases the chance of error in each k-mer. Another consideration is read overlap. If two reads overlap by less than k nucleotides, they do not share a vertex in the graph. This issue creates a coverage graph and decreases contig length. Consequently, the choice of k-mer length must be properly balanced for proper assembly. Musket was used to trim the raw sequence data and to optimize k-mer size[18] Musket is a multistage k-mer-based corrector for Illumina short read data. It utilizes the k-mer spectrum approach and introduces three different correction techniques in a multistage workflow, including two-sided conservative correction, one-sided aggressive correction, and voting-based refinement. Musket is multi-threaded using a master-slave model, so its parallel scalability outcompetes rival correctors. For many *de novo* assemblers, pre-assembly data filtering and cleaning is required, and Musket has successfully enhanced the assemblies of *Escherichia coli*, *Caenorhabditis elegans*, and Chr14 from the GAGE dataset[18]. However, some assembly programs like MaSuRCA recommend not cleaning the data before use.

Quality Assessment Programs

Sequence data must be assessed for quality before assembly. Poor sequence quality will result in inaccurate genome assembly and annotation, since the probability that each given base call is erroneous is high. Thus, any downstream analysis of incorrect input is less informative. Fundamentally, the quality of the raw sequence depends on the interaction of the sequencer with the specific genome of interest. Illumina next-generation sequencing includes Phred quality scores along with the sequence output[19]. This score indicates the probability that a given base was called incorrectly by the sequencer. While Phred scores are derived statistically from experimental sequencing tests, sequence quality and thus Phred scores are based on the sequencer's interpretation of the actual DNA. For example, high GC content in a nucleotide sequence is difficult for most sequencing platforms to interpret. Thus, the reported Phred score is low for that stretch of bases is low. However, the underlying cause, high GC content, is not reported. This information is contained in the sequencer's output, though, because the sequence is provided along with each base's Phred score.

FastQC provides an assessment of the quality of the raw sequencing data by simultaneously analyzing the provided Phred score and sequence[20]. FastQC is designed to spot issues either arising from the sequencer or the starting library material. FastQC first provides basic statistics like the total number of sequences, filtered sequences, and sequence length. FastQC primarily creates interactive graphs that reflect per base quality like the per base sequence quality, per base sequence content, per base GC content, and per base N content. FastQC also evaluates the per sequence quality scores, per sequence GC content, sequence length distribution, sequence duplication levels, overrepresented sequences, and K-mer content. FastQC is a simple way to quickly conduct quality control on raw sequence data from high throughput sequencing pipelines.

Once the sequence quality is ascertained, genome assembly can begin. Once a genome is assembled, it should be assessed for completeness and accuracy. As previously discussed, different genome assembly algorithms are optimized for different types of genomes. Codon usage bias and genomic GC content are species-specific, so genome assembly algorithms must be optimized for the species of interest to properly account for these nuances. Incorrect or low accuracy genome assembly causes poor gene identification and genome annotation. Thus, in a streamlined pipeline from genome sequencing to annotation, one step needs to be genome assembly assessment. BUSCO and QUAST were used to comparatively assess the qualities of the two assembled genomes.

BUSCO, Benchmarking Universal Single-Copy Orthologs, provides a quantitative assessment of genome assembly and annotation completeness based on evolutionarily informed expectations of gene content[21]. BUSCO provides major lineages with genes from orthologous groups present as single-copy orthologs in at least 90% of the species. This approach allows for rare duplications and losses while establishing an evolutionary informed expectation that the genes should be single-copy orthologs in the newly-sequenced species. BUSCO datasets currently represent 3023 genes for vertebrates, 2675 genes for arthropods, 843 genes for metazoans, 1438 genes for fungi, and 429 genes for eukaryotes. BUSCO has also adopted 40 universal marker genes for prokaryotic genomes. BUSCO utilizes HMMER, Blast+, and Augustus for its evolutionary insights. BUSCO assesses genome completeness by quantitating the numbers of complete BUSCO matches. BUSCO can also assess transcriptomes by incorporating EMBOSS transeq into the pipeline.

QUAST, Quality Assessment Tool for genome assemblies, is a popular genome assessment program[22]. It aggregates methods and quality metrics from existing software, and then it extends these programs with new metrics. QUAST utilizes E-MEM (an improvement over MUMmer), GeneMarkS, GeneMark-ES, GlimmerHMM, GAGE, and Gnuplot to define quality metrics for an assembly. QUAST can also find structural variants by utilizing a reference genome. However, a reference genome is not required to use QUAST. Bedtools is used to calculate raw and physical read coverage, which can be shown in Icarus contig alignment viewer. QUAST generates interactive plots for most of the measurements. QUAST can run on multi-core processors, thus increasing its efficiency and reducing its run time via parallelization.

Methods

Endophyte cultures originally isolated from *Panicum virgatum* were grown on potato dextrose agar plates for two weeks[10]. DNA extraction was performed using the 1000 Fungal Genomes Project protocol[24]. DNA sequencing was completed at the University of Texas at Austin through the Genome Sequencing and Analysis Facility. Paired-end sequencing (2x150) was conducted using the Illumina HiSeq 4000 system. The raw fastq files were trimmed and k-mer-corrected using Musket. Comparative genome assembly was completed using MaSuRCA, Velvet, and SOAPdenovo2. The recommendations given in each program's manual were followed for these genomes. MaSuRCA takes raw reads, so Musket was not used in the MaSuRCA assembly pipeline. MaSuRCA automatically selects its own k-mer value. Musket was used prior to the Velvet assemblies and some of the SOAPdenovo2 assemblies to test the dependence of assembly quality on input. K-mer values for these two program were selected with KmerGenie and SOAPdenovo2's recommendations. BUSCO and QUAST were separately utilized to assess the quality and completeness of the assembled genomes. Outputs from the genome assembly programs included Musket-corrected contigs and scaffold files as well as non-Musket-corrected contigs and scaffold files, which were differentially examined using BUSCO and QUAST. Many programs were run on the supercomputer Stampede at TACC, or the Texas Advanced Computing Center[25].

The annotation process is ongoing, so the results are not presented at this time. Annotation is based on the JGI Fungal Genome Annotation protocol recommendations[26] and incorporates previously collected transcriptome data based on the Broad Institute's recommendations[27].

Results

FastQC

The forward and reverse sequence reads for both *Cochliobolus kusanoi* and *Penicillium pinophilum* were high quality and required very little cleaning based on FastQC.

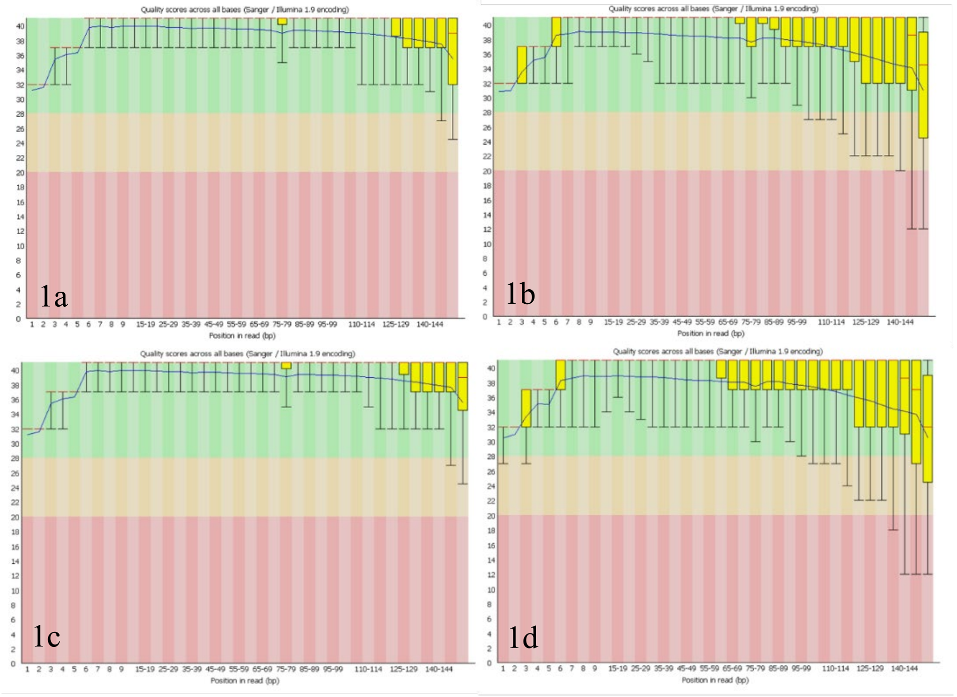


Figure 1**:** FastQC per base sequence quality results. The y-axis shows the quality scores encoded by the original Illumina raw fastq file. The higher the score, the better the base call. The y-axis is divided into very good quality (green), reasonable quality (orange), and poor quality (red). On most platforms, the quality degrades as the run progresses, so the falling quality at the tail end of the reads is not problematic. For each position, a box whisker plot is drawn. The central red line represents the median value. The blue line is the mean quality. The yellow box is the inter-quartile range. The upper and lower whiskers represent the 10% and 90% points. **1a:** Forward read quality for *C. kusanoi* **1b:** Reverse read quality for *C. kusanoi* **1c:** Forward read quality for *P. pinophilum* **1d:** Reverse read quality for *P. pinophilum*

The only significant failed module assessed by FastQC was for k-mer content. The error is issued when any k-mer is enriched more than 10-fold at any individual base position. Because no trimming was conducted before FastQC assessment, this was anticipated, as the adapters had not been removed from the raw sequencing data. Based on this quality reading, Musket was chosen to trim k-mers out of the raw sequence files.

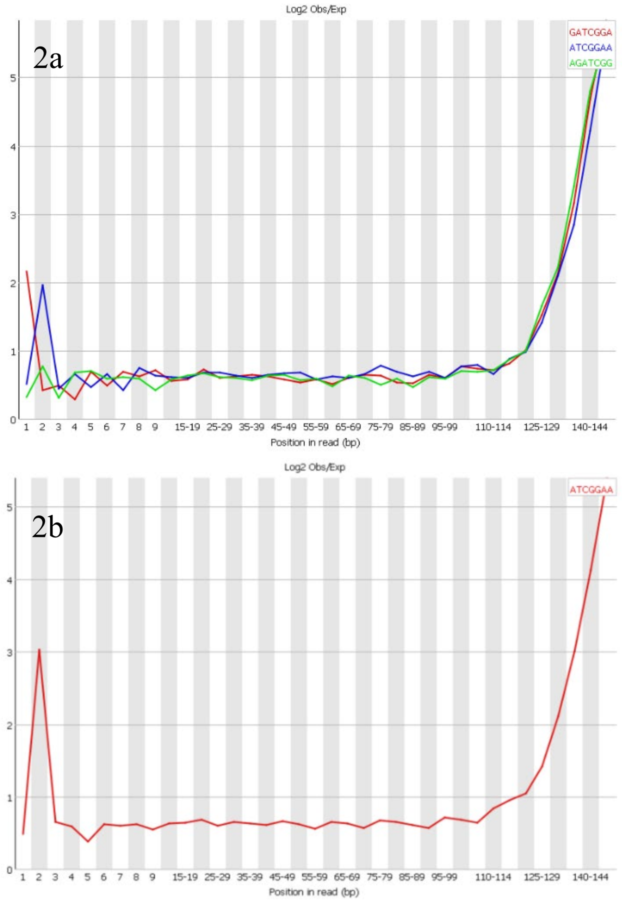


***Figure 2**:* K-mer enrichment is counted for every 7-mer within the sequence library. FastQC calculates an expected level of the k-mer based on the base content of the library as a whole. Then the program uses the actual count to calculate an observed/expected ratio for the k-mer. A pattern of bias was observed for k-mer content at different points over the read length. **2a:** K-mer content for *C. kusanoi* **2b:** K-mer content for *P. pinophilum*

Comparison of Genome Assemblies

Using QUAST, the cumulative length of each genome was determined. A reference genome of the same genus was selected from JGI for each fungus for comparison. The reference *Cochliobolus* genome for *C. kusanoi* differed more than the reference *Penicillium* genome for *P. pinophilum*. This is likely based on poor reference quality or a more phylogenetically distant reference for *Cochliobolus* compared to *Penicillium*.

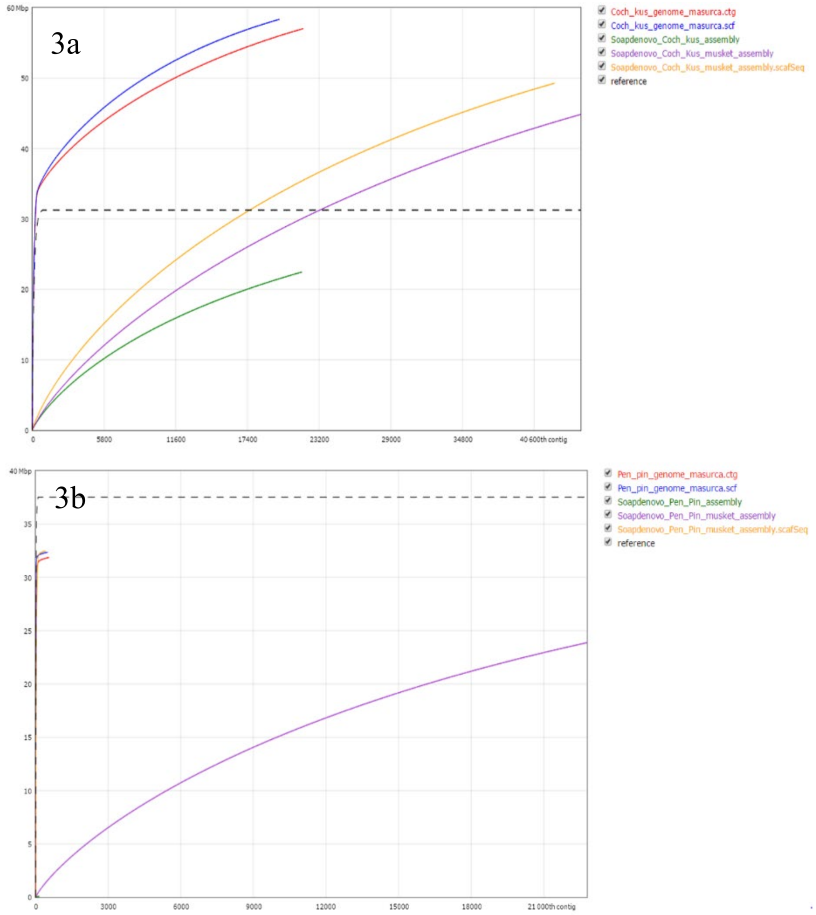


Figure 3: The cumulative length plots show the growth of contig lengths. On the x-axis, contigs are ordered the largest (contig #1) to the smallest. The y-axis gives the size of the x largest contigs in the assembly. The assembly type is color-coded. The MaSuRCA contigs genome assembly is red, the MaSuRCA scaffold genome assembly is blue, the SOAPdenovo2 contigs assembly is green, the SOAPdenovo2 Musket-corrected contigs assembly is purple, and the SOAPdenovo2 Musket-corrected scaffold assembly is orange. The reference sequence is the dotted black line. **3a:** While the reference *Cochliobolus* sequence was roughly 31 Mbp, most of the assemblies of *C. kusanoi* were larger. Most importantly, the MaSuRCA assemblies estimate the genome size to be about 57.5 Mbp. **3b:** The reference sequence based on a different *Penicillium* species was about 37 Mbp. Most of the *P. pinophilum* assemblies placed the genome to be about 32 Mbp.

QUAST was also used to determine the contig N length. N50 values are commonly used as an assessment of an assembled genome. N50 is the length for which the collection of all contigs of that length or longer covers at least half an assembly. N50 is thus partially dependent on the length of the genome. High N50 values relative to the size of the genome are desired. When a reference genome was provided to QUAST, the N50 value for *C. kusanoi* tripled while the N50 value for *P. pinophilum* decreased slightly.

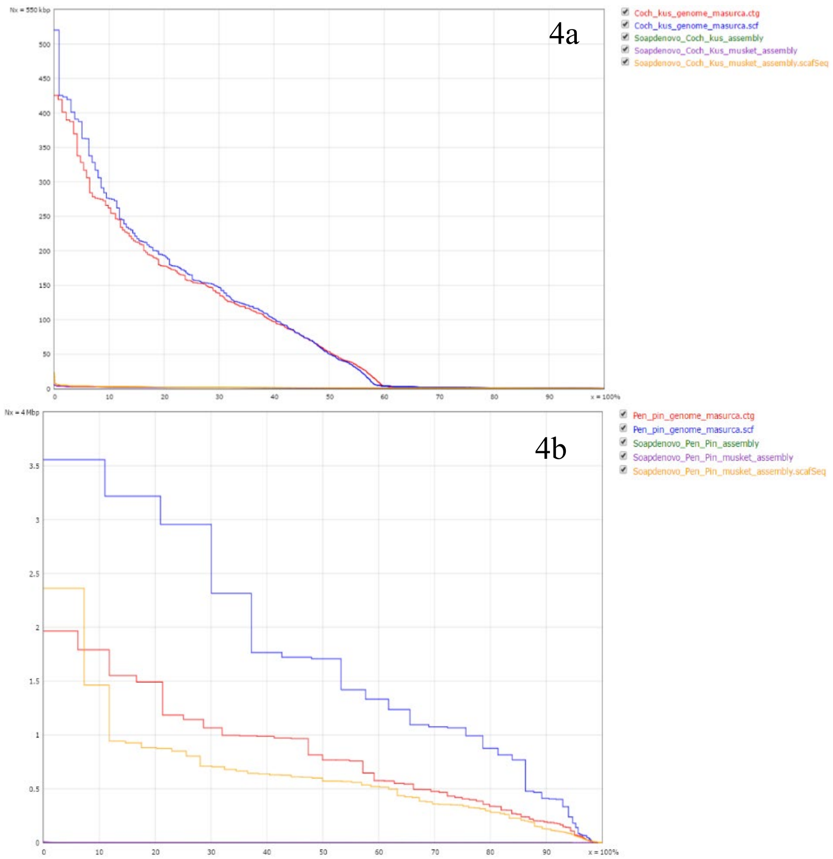


Figure 4: The Nx plots for *C. kusanoi* and *P. pinophilum* based the various genome assemblies. **4a:** The Nx values are presented for *C. kusanoi*. No reference genome was provided. The MaSuRCA assembly using contigs provided the best N50 value where N50 = 53,577 bp. **4b:** Here, the Nx values for *P. pinophilum* are plotted. No reference genome was used. The best N50 was produced by the MaSuRCA scaffold assembly with N50 = 1,708,132 bp.

Using BUSCO, the completeness of each genome assembly was assessed. BUSCO divides completeness metrics into complete and single copy, complete and duplicate copy, fragmented, and missing assessments.

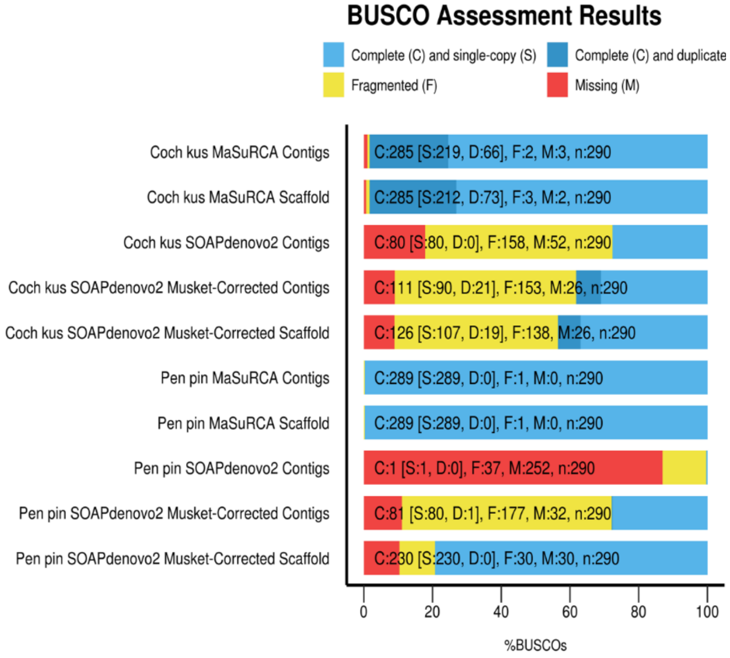


Figure 5: The MaSuRCA contigs and scaffold assemblies performed equally well and by far the best when assessed by BUSCO for *C. kusanoi* and *P. pinophilum*. SOAPdenovo2 assemblies improved significantly when Musket was used to trim the sequences before assembly. SOAPdenovo2 assessments were further improved when the output assemblies were scaffolds, not contigs.

Discussion

Based on BUSCO and QUAST assessments, the de novo MaSuRCA genome assemblies for both *Cochliobolus kusanoi* and *Penicillium pinophilum* outperformed SOAPdenovo2 based on BUSCO and N50 metrics. BUSCO's results showed that MaSuRCA generated a 98.3% complete genome assembly for *C. kusanoi* and a 99.7% complete genome assembly for *P. pinophilum*. For *C. kusanoi*, SOAPdenovo2 produced 27.6%, 38.2%, and 43.5% genome completeness using contigs, Musket-corrected contigs, and the Musket-corrected scaffold, respectively. The genome assembly for *C. kusanoi* from MaSuRCA is more than twice as complete as the best SOAPdenovo2 assembly based on BUSCO assessments. For *P. pinophilum,* MaSuRCA generated a 99.7% complete genome using either contigs or a scaffold. For comparison, SOAPdenovo2 produced 0.3%, 27.9%, and 79.3% genome completeness using contigs, Musket-corrected contigs, and the Musket-corrected scaffold, respectively. The genome assembly produced by MaSuRCA for *P. pinophilum* was over 20% more complete than the best SOAPdenovo2 assembly.

Using QUAST, the *de novo* assemblies were assessed using N50 values. For *C. kusanoi,* the contigs and scaffold from MaSuRCA produced N50 values of 53,577 and 50,815 bases, respectively. The contigs, Musket-corrected contigs, and Musket-corrected scaffold from SOAPdenovo2 has N50 values of 1125, 1089, and 1313 bases, respectively. For *P. pinophilum,* the contigs and scaffold from MaSuRCA produced N50 values of 768,585 and 1,708,132 bases, respectively. The contigs, Musket-corrected contigs, and Musket-corrected scaffold from SOAPdenovo2 has N50 values of 563, 1141, and 571,787 bases, respectively.

N50 values are dependent upon the length of the genome. Because the presented assemblies are the first for each fungus, N50 value comparisons to other published fungal genomes are not particularly informative. However, BUSCO assessments were completed on three JGI *Cochliobolus* genome assemblies and three JGI *Penicillium* genome assemblies for quality reference[28,29,30,31,32,33]. On average, the published *Cochliobolus* genomes had a BUSCO completeness score of 98.3% with 0.3% fragments and 1.4% missing. This is on par with the *C. kusanoi* assemblies produced by MaSuRCA, which had 98.3% completeness, 1.0% fragments, and 0.7% missing. On average, the three tested *Penicillium* species had a completeness score of 99.1% with 0.1% fragments and 0.8% missing. The completed *P. pinophilum* produced by MaSuRCA had a marginally improved completeness of 99.7% with 0.3% fragments and 0.0% missing. The quality of the MaSuRCA assembler for these genomes is comparable to the high-quality genomes published by the JGI.

Based on these assessments, MaSuRCA should be used for future fungal genome assemblies. These are the first genome assemblies for each fungus. Because these assemblies are accurate and compete, they are suitable for annotation and other downstream analyses.

Sequence trimming and scaffolding differentially influenced the final genome assemblies depending on the assembly program, though MaSuRCA required unfiltered and untrimmed input sequences, and MaSuRCA performs approximately as well using either the contigs or scaffold files**.** Scaffolding algorithms vary depending on the assembly program. SOAPdenovo2 constructs scaffolds by utilizing paired-end reads beginning with short insert sizes followed iteratively to large insert sizes. Novel approaches to heterozygous contig pair detection, chimeric scaffold construction, and contig matching with insufficient paired-end information were implemented in

SOAPdenovo2 compared to SOAPdenovo[16]. These added mechanisms to improve contig relationships and untangle chimeras likely improved the evolutionary completeness and contig length seen in the BUSCO and QUAST results. MaSuRCA is a modified version of the CABOG assembler[13]. For scaffolding, CABOG primarily uses mate-pair information (https://academic.oup.com/bioinformatics/article/24/24/2818/197033/Aggressive-assembly-of-pyrosequencing-reads-with). However, in this project, mate-pairs were not generated. Consequently, the scaffold did not show improvement over the contigs in MaSuRCA output assessment. Reference genome incorporation improved the N50 value for *C. kusanoi* but not *P. pinophilum*. This is likely due to the relative phylogenetic distance between the fungus of interest and the available reference sequence.

The quality of Velvet's genome assemblies will be assessed. However, preliminary results suggest the assemblies by Velvet are better than SOAPdenovo2 but not as good as MaSuRCA. Thus, Velvet will not be investigated too much further for our annotation analysis, since the MaSuRCA assemblies are high quality. The ability of Scaffold_Builder to enhance genome completeness should be assessed, particularly if Velvet or SOAPdenovo2 are to be used for assembly and a reference sequence is available. Genome annotation work is still in its early stages and will provide deeper genomic insight once completed. Leveraging transcriptome data allows for better genome annotation. Differentially expressed genes between the two selected fungal species will indicate potential genes responsible for the mutualist or antagonist associations with the host plant. This pipeline, once finished, will be streamlined for fungal genomic analysis, enabling the quick assembly, assessment, and annotation of future genomes. This genetic analysis of fungal phenotypes is a new, unique approach to understanding fungi-host interactions, particularly for endophytes.

Genetic understanding of *C. kusanoi* contrasted to *P. pinophilum* will grow as annotation is completed, so the intricacies of each fungus's behavior can be uncovered and characterized. Fungal genome annotation provides insight into both genetics and applied genetics, including mechanisms of fungal genome evolution, fungi-specific gene family innovations, genomic potential for sexual cycles, functional genomics, structural and functional characterization of protein-coding and non-coding genes, and other genomic features. Aggressive pathogenic fungi tend to possess a set of pathogenicity genomic signatures[34]. Similar regions are expected to be differentially expressed in the antagonist versus the mutualist fungus. The potential regulatory patterns for gene clusters can also be evaluated during the annotation phase. Screening of the genome by the Antibiotics & Secondary Metabolite Analysis Shell (AntiSMASH) pipeline will indicate possible secondary metabolites produced by these fungi (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489286/). The fungi likely produce many more metabolites than AntiSMASH will hypothesize since many metabolites have yet to be identified, particularly in fungi. However, this rudimentary metabolite assessment is a good starting point for any future experimental metabolic studies. This methodology can be applied to other disparate sets of fungi to relate genomic underpinnings to morphology and symbioses.

This is the first comparison of genome assembly programs optimized for fungi. Other popular assemblers, such as ABySS, DISCOVAR, or SSAKE, can be tested in the future. However, given the limitations of our sequencing platform, only certain assembly programs could be tested. For example, ALLPATHS-LG, utilized by JGI, requires two paired-end libraries, one short and one long[35]. This assembly method was not possible given the sequencing choice for this project. Sequencing platforms and coverage should be strategically selected based on the range of assembly options desired.

References

1. Schatz, M.C., A.L. Delcher, and S.L. Salzberg. 2010. Assembly of large genomes using second-generation sequencing. *Genome Res* 20(9):1165-73. doi: 10.1101/gr.101360.109

2. Bacon, C., and White, J. (Eds.) *Microbial Endophytes*. Marcel Dekker, Inc., 2000. New York, NY, pp 3-29.

3. Elbersen HW, West CP (1996) Growth and water relations of field-grown tall fescue as influenced by drought and endophyte. *Grass and Forage Science* 51(4):333–342. doi:10.1111/j.1365-2494.1996.tb02068.x

4. Malinowski D. P. and D. B. Belesky. 2000. Adaptations of endophyte-infected cool-season grasses to environmental stresses: Mechanisms of drought and mineral stress tolerance. *Crop Science* 40: 923–940.

5. Waller F., B. Achatz, H. Baltruschat, J. Fodor, K. Becker, M. Fischer, T. Heier, et al. 2005. The endophytic fungus Piriformospora indica reprograms barley to salt-stress tolerance, disease resistance, and higher yield. *Proceedings of the National Academy of Sciences USA* 102: 13386–13391.

6. Rodriguez R. J., J. F. White Jr., A. E. Arnold, and R. S. Redman. 2009. Fungal endophytes: Diversity and functional roles. *New Phytologist* 182: 314–330.

7. Rodriguez R. J., C. Woodward, and R. S. Redman. 2010. Adaptation and survival of plants in high stress habitats via fungal endophyte conferred stress tolerance. *In* J. Seckbach and M. Grube [eds.], Symbioses and Stress, vol. 17, Cellular origin, life in extreme habitats and astrobiology, 463–476. Springer, Dordrecht, Netherlands.

8. Morsy M. R., J. Oswald, J. He, Y. Tang, and M. J. Roossinck. 2010. Teasing apart a three-way symbiosis: Transcriptome analyses of *Curvularia protuberata* in response to viral infection and heat stress. *Biochemical and Biophysical Research.*

9. Worchel, E.R., Giauque, H.E. & Kivlin, S.N. Fungal symbionts alter plant drought response. *Microb Ecol* (2013) 65:671. doi:10.1007/s00248-012-0151-6

10. Giauque, H., & Hawkes, C.V. Climate affects symbiotic fungal endophyte diversity and performance. *American Journal of Botany* (2013) 100:7. doi: 10.3732/ajb.1200568.

11. Giauque, H., & Hawkes, C.V. Historical and current climate drive spatial and temporal patterns in fungal endophyte diversity. *Fungal Ecology* (2016) 20: 108-114. doi:10.1016/j.funeco.2015.12.005.

12. Manamgoda, D.S., Cai, L., McKenzie, E.H.C. et al. Fungal Diversity (2012) 56: 131. doi:10.1007/s13225-012-0189-2

13. Aleksey V. Zimin, Guillaume Marçais, Daniela Puiu, Michael Roberts, Steven L. Salzberg, James A. Yorke; The MaSuRCA genome assembler. *Bioinformatics* 2013; 29 (21): 2669-2677. doi: 10.1093/bioinformatics/btt476

14. Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. An Eulerian path approach to DNA fragment assembly. *PNAS* 2001, 98 (17) 9748-975. doi:10.1073/pnas.171285098

15. Zerbino, Daniel R., and Ewan Birney. Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Research* 18.5 (2008): 821–829. *PMC*. Web. 26 Apr. 2017.

16. Luo, Ruibang et al. SOAPdenovo2: An Empirically Improved Memory-Efficient Short-Read *de Novo* Assembler. *GigaScience* 1 (2012): 18. *PMC*. Web. 26 Apr. 2017.

17. Rayan Chikhi, Paul Medvedev; Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 2014; 30 (1): 31-37. doi:10.1093/bioinformatics/btt310.

18. Yongchao Liu, Jan Schröder, Bertil Schmidt; Musket: a multistage *k*-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* 2013; 29 (3): 308-315. doi: 10.1093/bioinformatics/bts690

19. "Quality scores for next-generation sequencing." Sequencing Literature. Illumina, 2017. Web. 02 May 2017. https://support.illumina.com/sequencing/literature.html.

20. "FastQC." *Babraham Bioinformatics*. Babraham Bioinformatics, 3 Aug. 2016. Web. 26 Apr. 2017, http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

21. Felipe A. Simão, Robert M. Waterhouse, Panagiotis Ioannidis, Evgenia V. Kriventseva, Evgeny M. Zdobnov. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015; 31 (19): 3210-3212. doi:10.1093/bioinformatics/btv351

22. Gurevich, Alexey et al. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* 29.8 (2013): 1072–1075. *PMC*. Web. 26 Apr. 2017. doi:10.1093/bioinformatics/btt086

23. Silva et al. Combining de novo and reference-guided assembly with scaffold_builder. *Source Code for Biology and Medicine* (2013) 8:23. doi:10.1186/1751-0473-8-23

24. Kohler, A., and Martin, F. “Genomic DNA Extraction.” *1000 Fungal Genomes Project*. 17 Jan. 2011, http://1000.fungalgenomes.org/home/protocols/high-quality-genomic-dna-extraction/.

25. "Stampede." *Texas Advanced Computing Center*. University of Texas at Austin, 2017. Web. 26 Apr. 2017, https://www.tacc.utexas.edu/stampede/.

26. *MycoCosm*. DOE Joint Genome Institute, 2017. Web. 26 Apr. 2017, http://jgi.doe.gov/data-and-tools/mycocosm/.

27. "Fungal Genomics." Fungal Genome Initiative. Broad Institute, 2017. Web. 26 Apr. 2017, https://www.broadinstitute.org/fungal-genome-initiative.

28. Turgeon, Gillian. "Cochliobolus carbonum 26-R-13." *JGI Genome Portal*. DOE Joint Genome Institute, 2012. Web. 02 May 2017.

29. Baker, Scott. “Cochliobolus heterostrophus C4 Standard Draft.” *JGI Genome Portal*. DOE Joint Genome Institute, 2011. Web. 02 May 2017.

30. Turgeon, Gillian. “Cochliobolus miyabeanus ATCC 44560." *JGI Genome Portal*. DOE Joint Genome Institute, 2012. Web. 02 May 2017.

31. Greenshields, Dave. “Penicillium bilaiae, ATCC 20851 Annotated Standard Draft.” *JGI Genome Portal*. DOE Joint Genome Institute, 2013. Web. 02 May 2017.

32. Greenshields, Dave. “Penicillium aculeatum, ATCC 10409 Annotated Standard Draft.” *JGI Genome Portal*. DOE Joint Genome Institute, 2014. Web. 02 May 2017.

33. Greenshields, Dave. “Penicillium brevicompactum AgRF18, Annotated Standard Draft.” *JGI Genome Portal*. DOE Joint Genome Institute, 2014. Web. 02 May 2017.

34. Kämper, Jörg, Regine Kahmann, Michael Bölker, Li-Jun Ma, Thomas Brefort, Barry J. Saville, Flora Banuett, James W. Kronstad, Scott E. Gold, Olaf Müller, Michael H. Perlin, Han A. B. Wösten, Ronald De Vries, José Ruiz-Herrera, Cristina G. Reynaga-Peña, Karen Snetselaar, Michael McCann, José Pérez-Martín, Michael Feldbrügge, Christoph W. Basse, Gero Steinberg, Jose I. Ibeas, William Holloman, Plinio Guzman, Mark Farman, Jason E. Stajich, Rafael Sentandreu, Juan M. González-Prieto, John C. Kennell, Lazaro Molina, Jan Schirawski, Artemio Mendoza-Mendoza, Doris Greilinger, Karin Münch, Nicole Rössel, Mario Scherer, Miroslav Vraneš, Oliver Ladendorf, Volker Vincon, Uta Fuchs, Björn Sandrock, Shaowu Meng, Eric C. H. Ho, Matt J. Cahill, Kylie J. Boyce, Jana Klose, Steven J. Klosterman, Heine J. Deelstra, Lucila Ortiz-Castellanos, Weixi Li, Patricia Sanchez-Alonso, Peter H. Schreier, Isolde Häuser-Hahn, Martin Vaupel, Edda Koopmann, Gabi Friedrich, Hartmut Voss, Thomas Schlüter, Jonathan Margolis, Darren Platt, Candace Swimmer, Andreas Gnirke, Feng Chen, Valentina Vysotskaia, Gertrud Mannhaupt, Ulrich Güldener, Martin Münsterkötter, Dirk Haase, Matthias Oesterheld, Hans-Werner Mewes, Evan W. Mauceli, David DeCaprio, Claire M. Wade, Jonathan Butler, Sarah Young, David B. Jaffe, Sarah Calvo, Chad Nusbaum, James Galagan, and Bruce W. Birren. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature*, 2006.444:97 – 101. doi:10.1038/nature05248.

35. Sante Gnerre, Iain MacCallum, Dariusz Przybylski, Filipe J. Ribeiro, Joshua N. Burton, Bruce J. Walker, Ted Sharpe, Giles Hall, Terrance P. Shea, Sean Sykes, Aaron M. Berlin, Daniel Aird, Maura Costello, Riza Daza, Louise Williams, Robert Nicol, Andreas Gnirke, Chad Nusbaum, Eric S. Lander, and David B. Jaffe. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS* 2011; 108 (4): 1513-1518; published ahead of print December 27, 2010, doi:10.1073/pnas.1017351108

Supplement of Enlarged Figures

Figure 1a

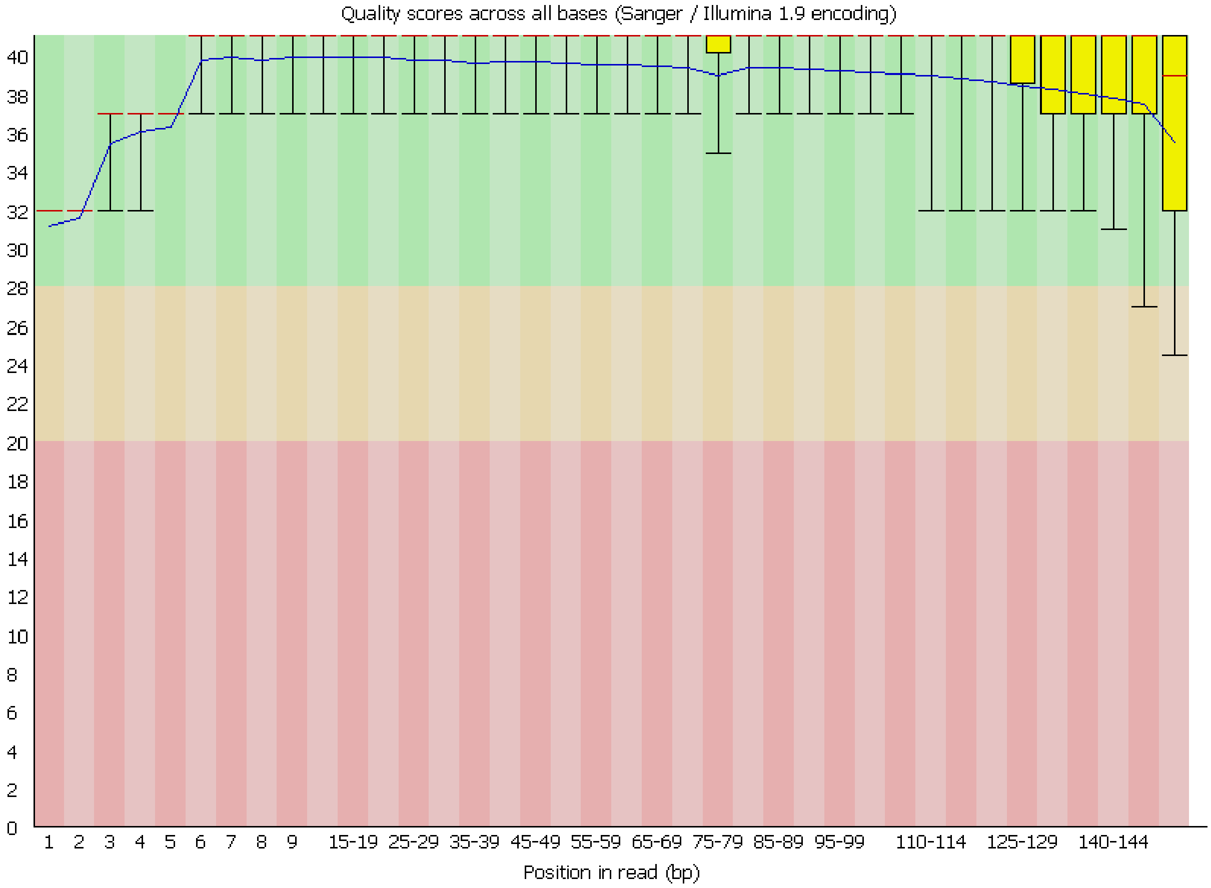


Figure 1b

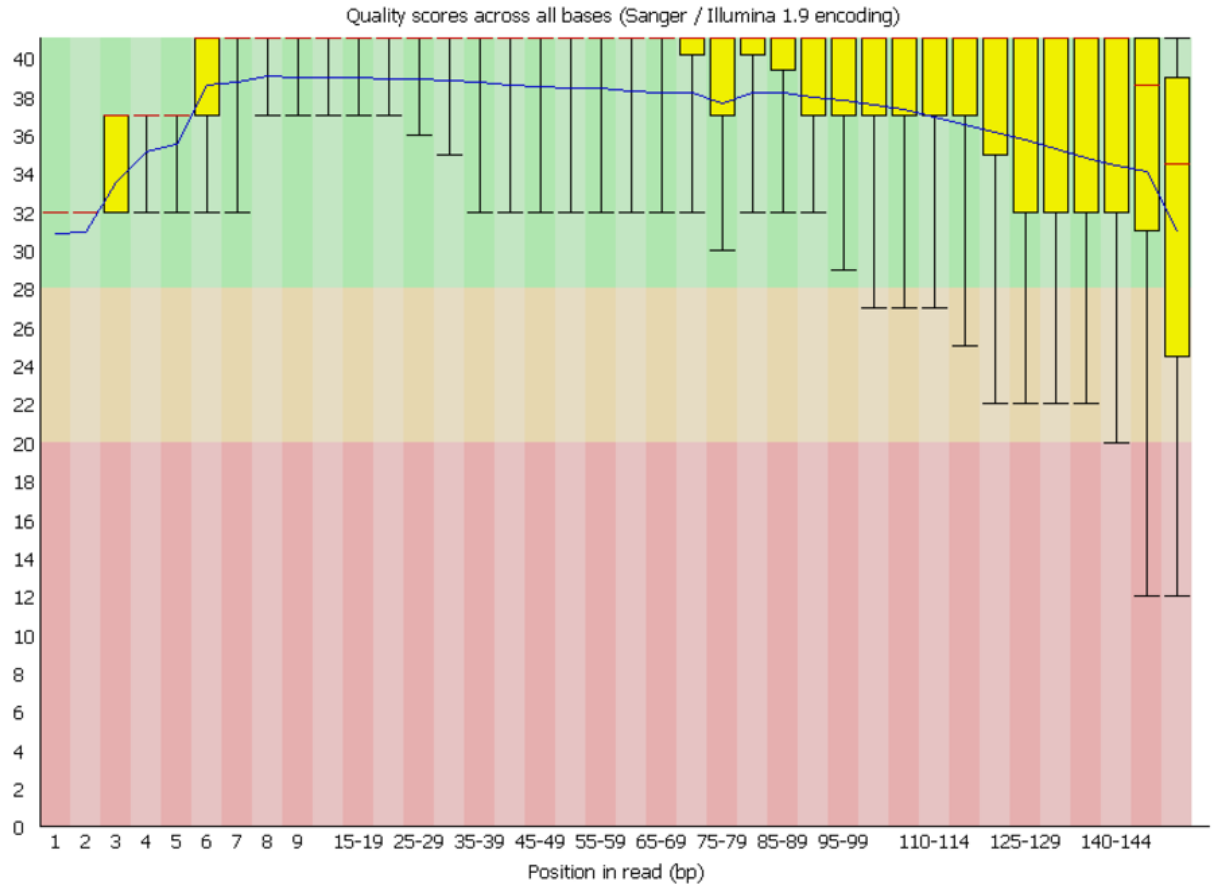

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
110-114
125-129
140-144
Position in read (bp)

Figure 1c

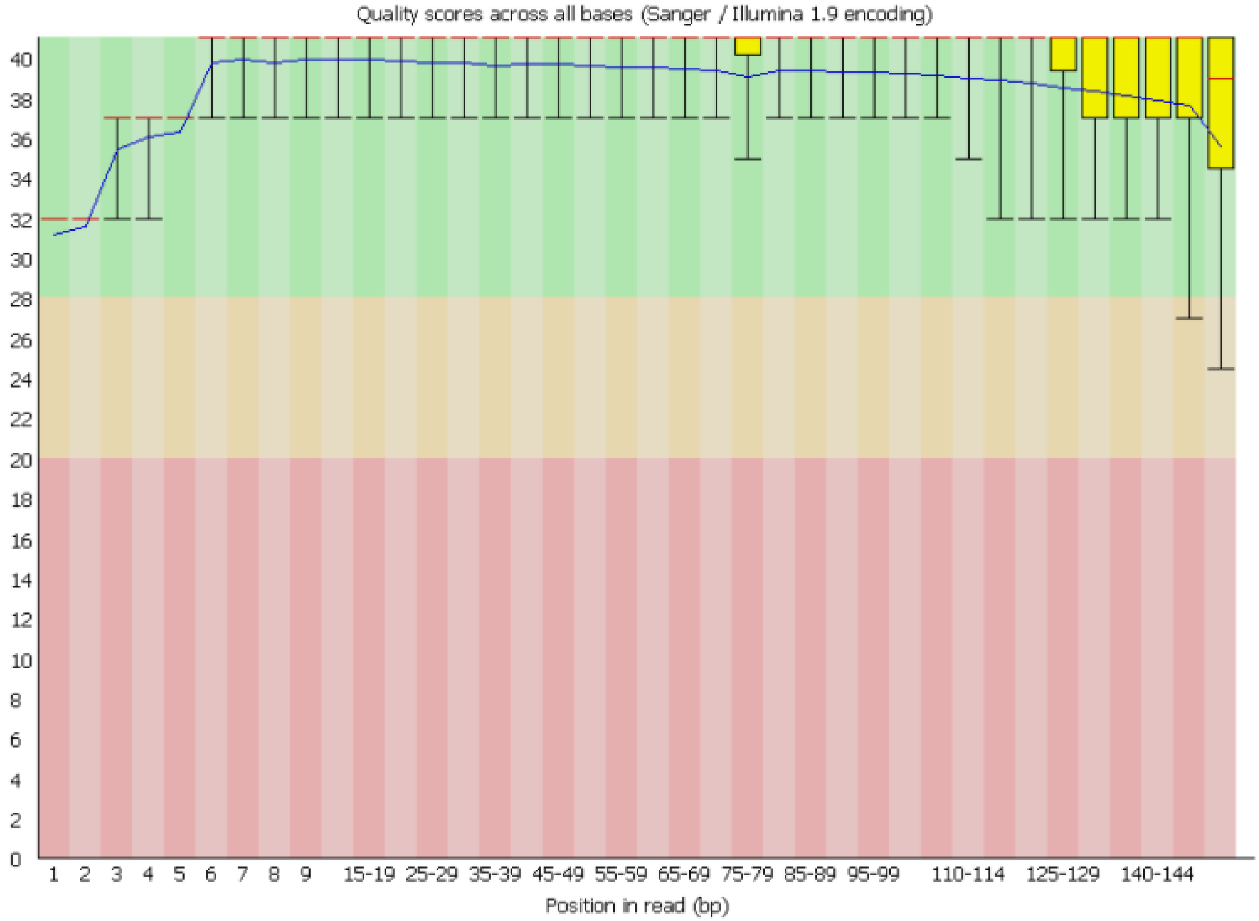

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
110-114
125-129
140-144
Position in read (bp)

Figure 1d

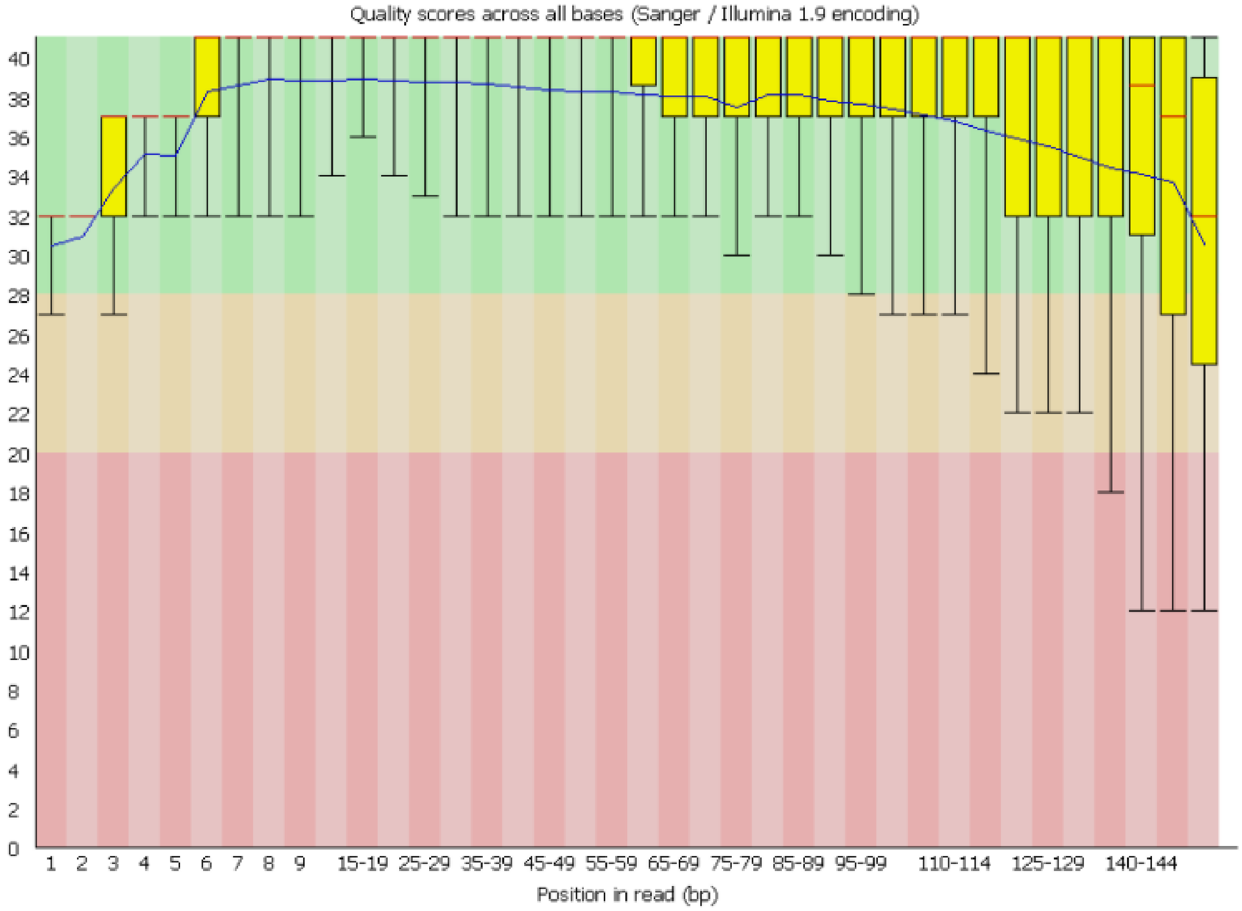

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
110-114
125-129
140-144
Position in read (bp)

Figure 2a

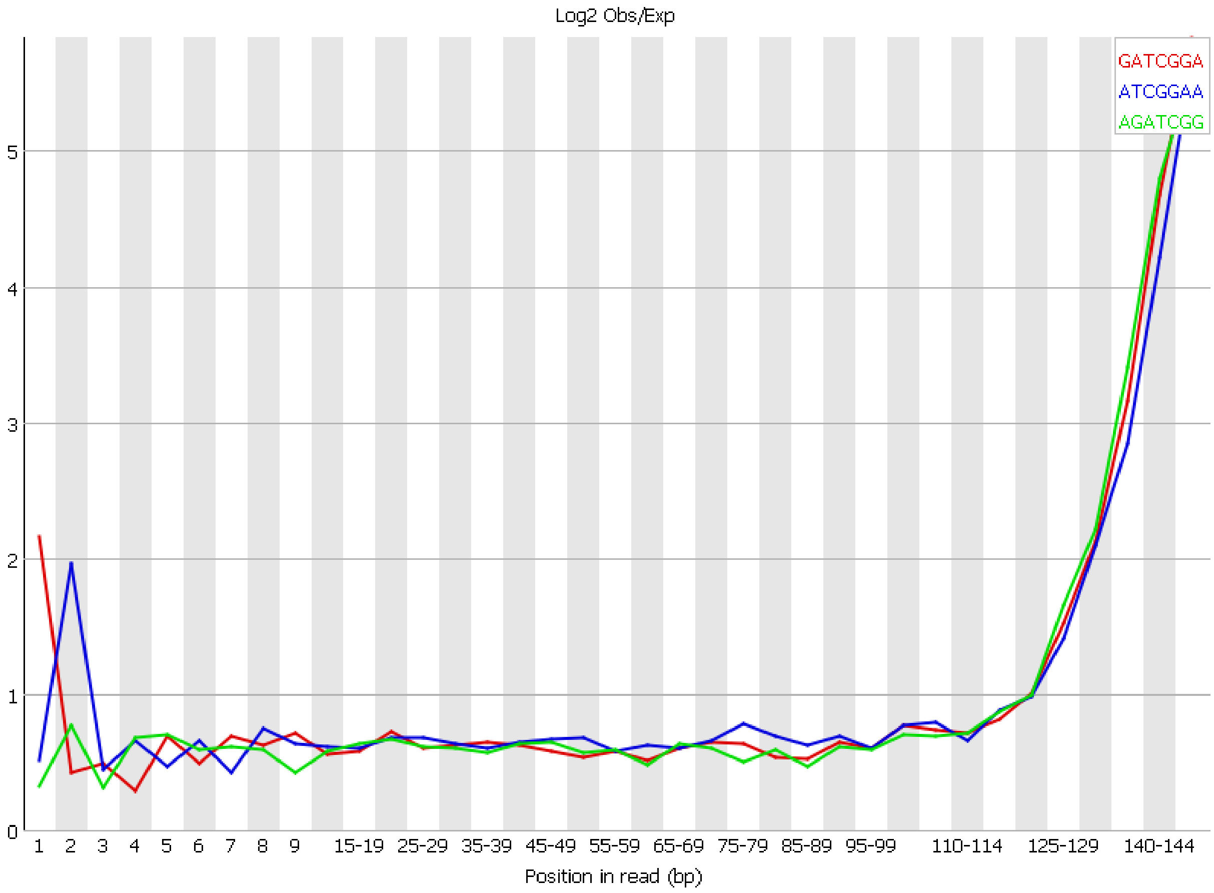

Log2 Obs/Exp
GATCGGA
ATCGGAA
AGATCGG
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
110-114
125-129
140-144
Position in read (bp)

Figure 2b

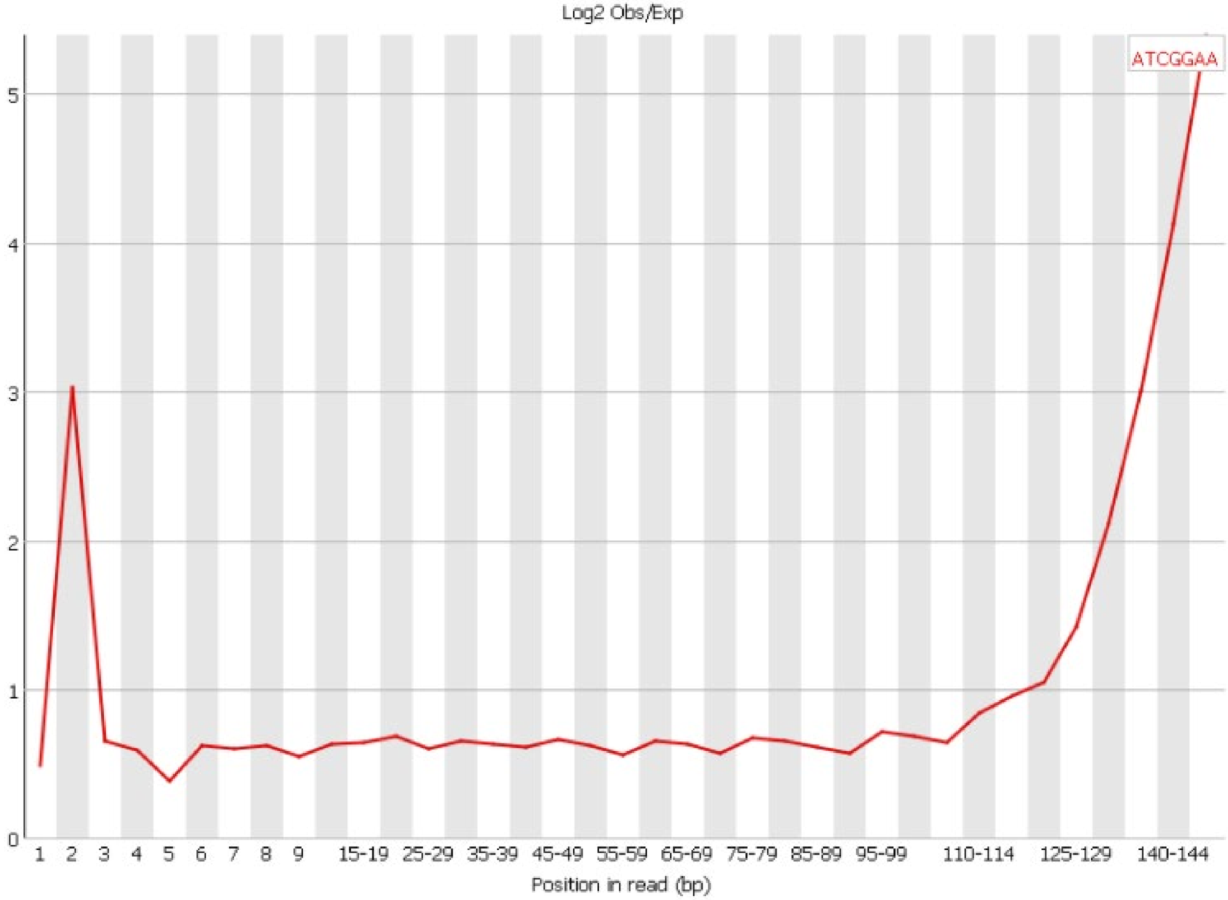

Log2 Obs/Exp
ATCGGAA
0
1
2
3
4
5
1
2
3
4
5
6
7
8
9
15-19
25-29
35-39
45-49
55-59
65-69
75-79
85-89
95-99
110-114
125-129
140-144
Position in read (bp)

Figure 3a

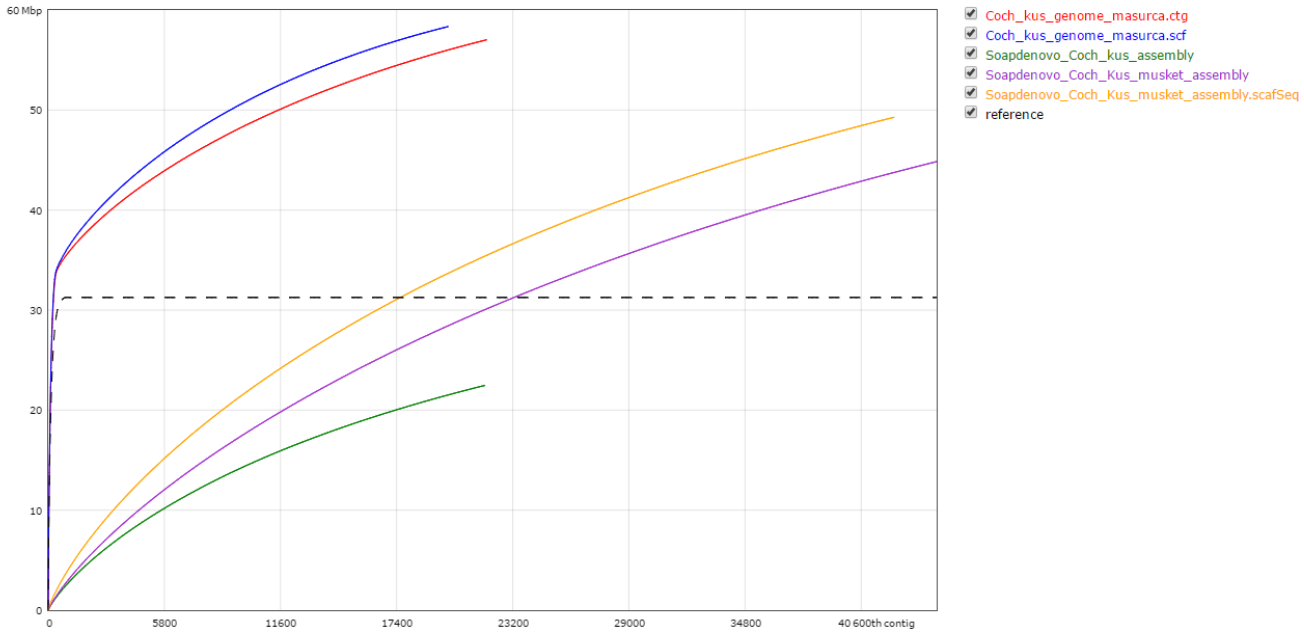


Figure 3b

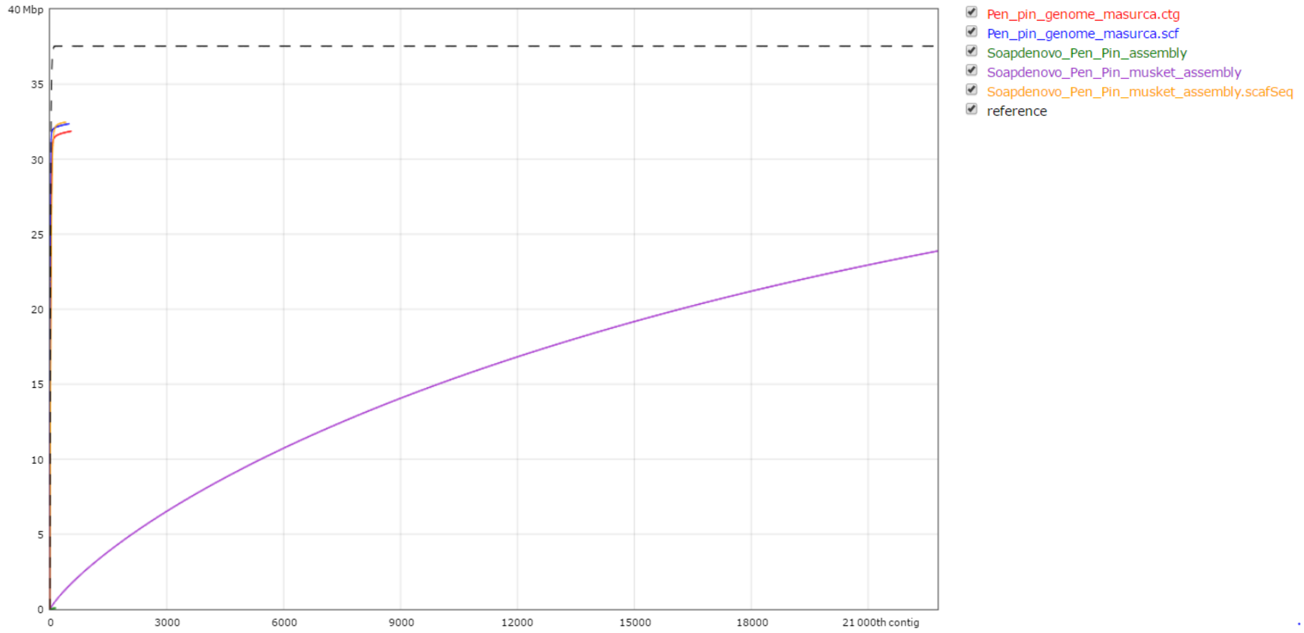


Figure 4a

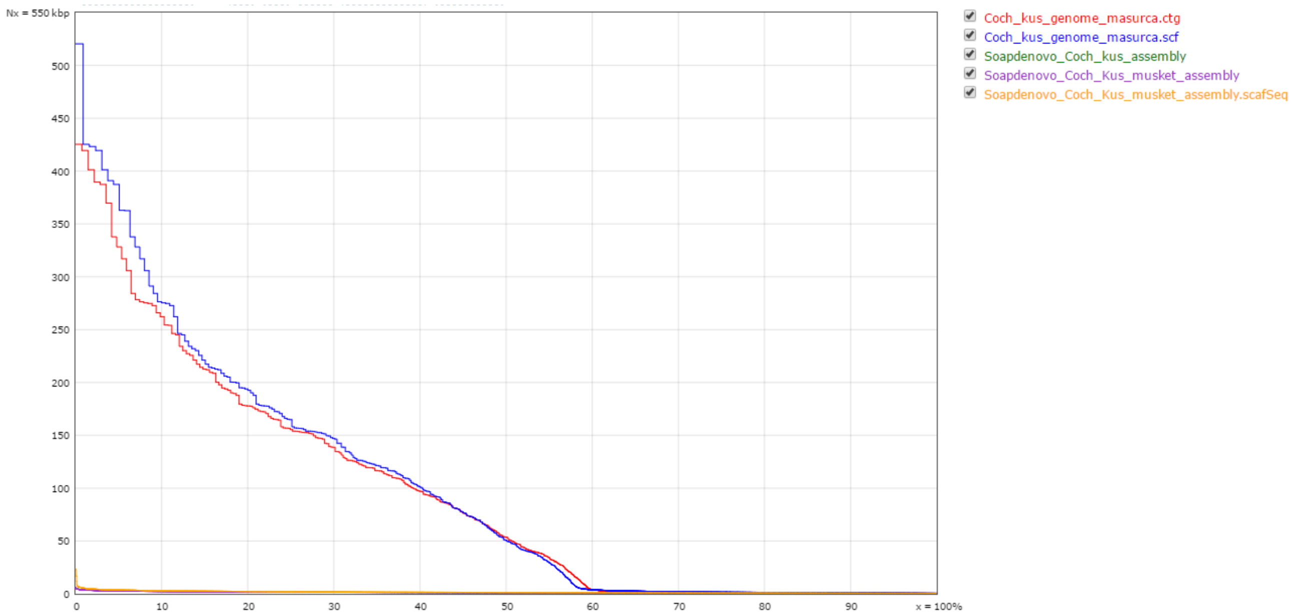


Figure 4b

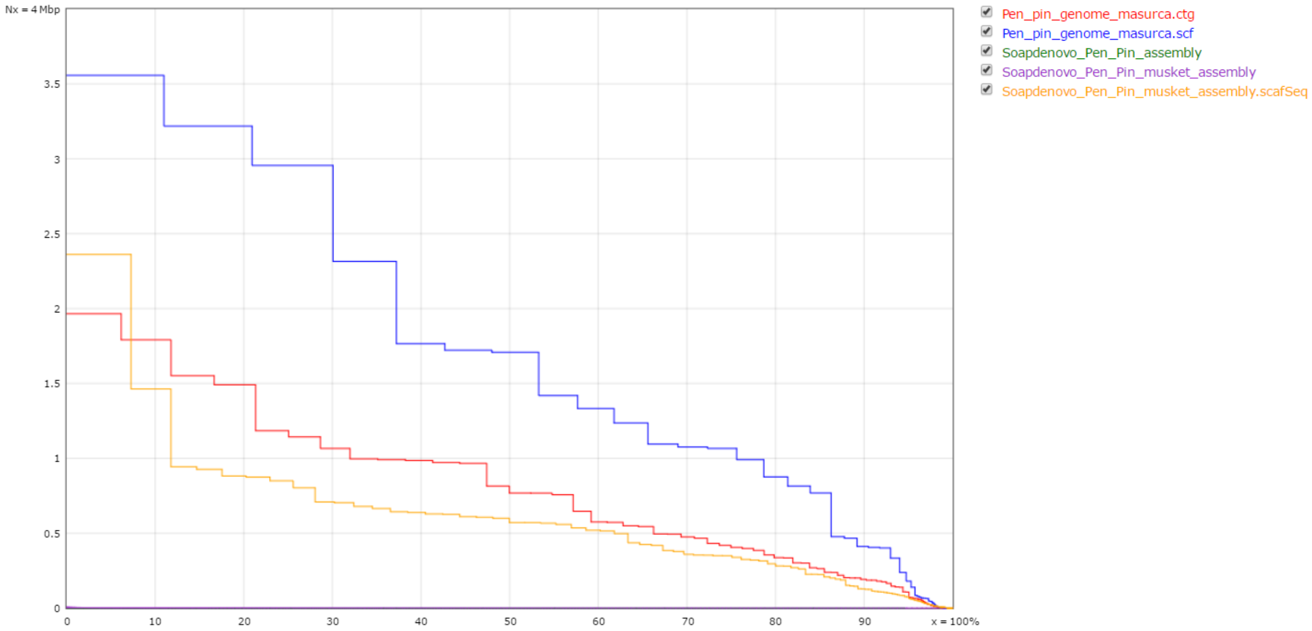
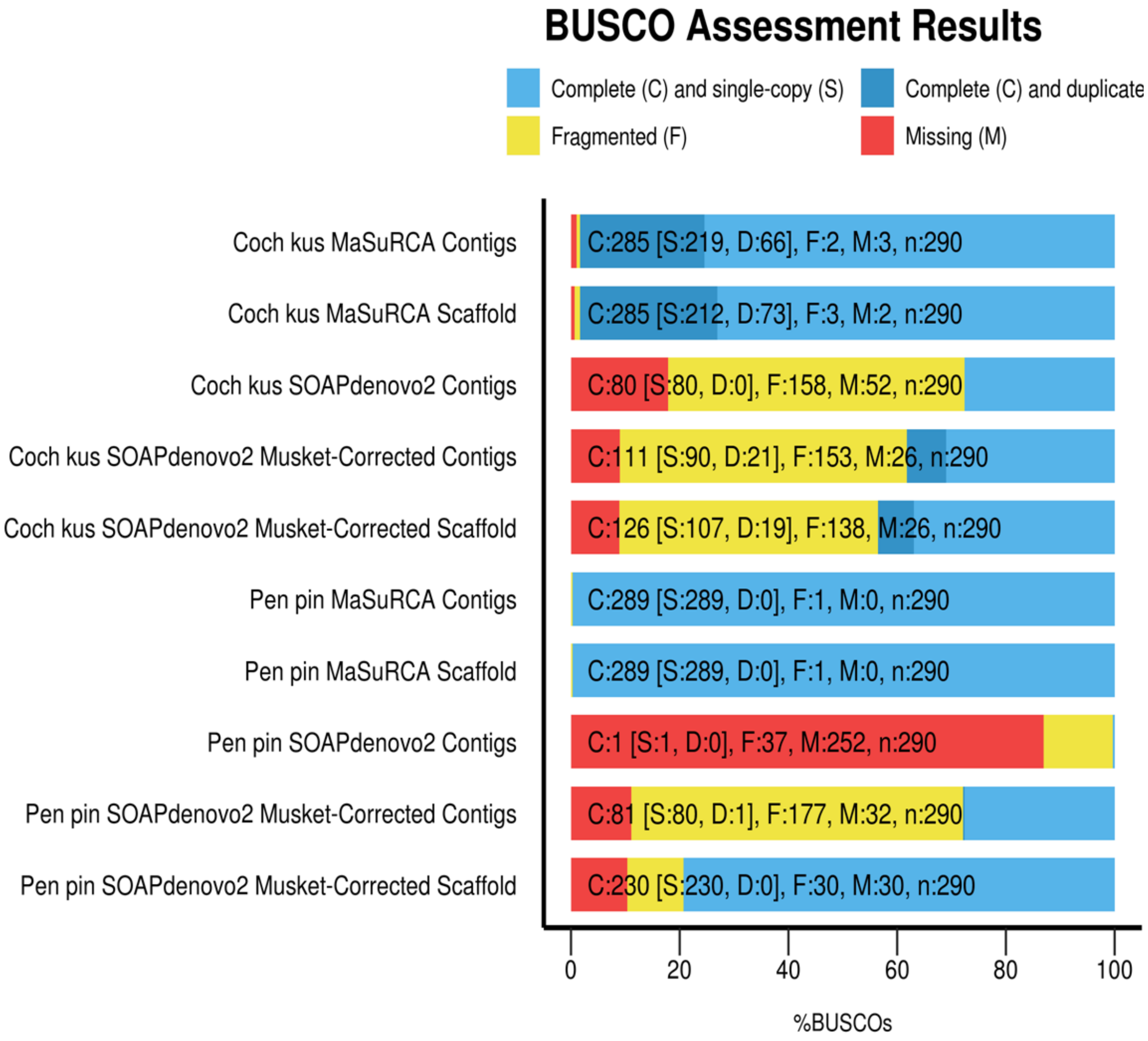


Figure 5


BUSCO Assessment Results
Complete (C) and single-copy (S)
Complete (C) and duplicate
Fragmented (F)
Missing (M)
Coch kus MaSuRCA Contigs
Coch kus MaSuRCA Scaffold
Coch kus SOAPdenovo2 Contigs
Coch kus SOAPdenovo2 Musket-Corrected Contigs
Coch kus SOAPdenovo2 Musket-Corrected Scaffold
Pen pin MaSuRCA Contigs
Pen pin MaSuRCA Scaffold
Pen pin SOAPdenovo2 Contigs
Pen pin SOAPdenovo2 Musket-Corrected Contigs
Pen pin SOAPdenovo2 Musket-Corrected Scaffold
C:285 [S:219, D:66], F:2, M:3, n:290
C:285 [S:212, D:73], F:3, M:2, n:290
C:80 [S:80, D:0], F:158, M:52, n:290
C:111 [S:90, D:21], F:153, M:26, n:290
C:126 [S:107, D:19], F:138, M:26, n:290
C:289 [S:289, D:0], F:1, M:0, n:290
C:289 [S:289, D:0], F:1, M:0, n:290
C:1 [S:1, D:0], F:37, M:252, n:290
C:81 [S:80, D:1], F:177, M:32, n:290
C:230 [S:230, D:0], F:30, M:30, n:290
0
20
40
60
80
100
%BUSCOs